WSSA
WEED SCIENCE SOCIETY OF AMERICA

Weed Vegetation of Sugarcane Cropping Systems of Northern Argentina: Data-Mining Methods for Assessing the Environmental and Management Effects on Species Composition

D. O. Ferraro, C. M. Ghersa, and D. E. Rivero*

Weed composition may vary because of natural environment, management practices, and their interactions. In this study we presented a systematic approach for analyzing the relative importance of environmental and management factors on weed composition of the most conspicuous species in sugarcane. A data-mining approach represented by *k*-means cluster and classification and regression trees (CART) were used for analyzing the 11 most frequent weeds recorded in sugarcane cropping systems of northern Argentina. Data of weed abundance and explanatory factors contained records from 1976 sugarcane fields over 2 consecutive years. The *k*-means method selected five different weed clusters. One cluster contained 44% of the data and exhibited the lowest overall weed abundance. The other four clusters were dominated by three perennial species, bermudagrass, johnsongrass, and purple nutsedge, and the annual itchgrass. The CART model was able to explain 44% of the sugarcane's weed composition variability. Four of the five clusters were represented in the terminal nodes of the final CART model. Sugarcane burning before harvesting was the first factor selected in the CART, and all nodes resulting from this split were characterized by low abundance of weeds. Regarding the predictive power of the variables, rainfall and the genotype identity were the most important predictors. These results have management implications as they indicate that the genotype identity would be a more important factor than crop age when designing sugarcane weed management. Moreover, the abiotic control of crop–weed interaction would be more related to rainfall than the environmental heterogeneity related to soil type, for example soil fertility. Although all these exploratory patterns resulting from the CART data-mining procedure should be refined, it became clear that this information may be used to develop an experimental framework to study the factors driving weed assembly.

Nomenclature: Bermudagrass, *Cynodon dactylon* Pers. (CYNDA); johnsongrass, *Sorghum halepense* (L.) Pers. (SORHA); purple nutsedge, *Cyperus rotundus* L. (CYPRO); itchgrass, *Rottboellia exaltata* (L.) L.f.(ROOEX).
Key words: Sugarcane, weed composition, classification and regression trees, statistics.

Predicting changes in weed species composition in field crops requires an understanding of the effect of management and biophysical factors on cropping systems. Data about the nature of these influences should help to design efficient weed-management regimes in cropping systems (Martínez-Ghersa et al. 2000). The link between explanatory factors and weed species composition is usually explored through multivariate direct gradient analyses for examining the relationships between factor and response variables measured on the same sampling units (Firehun and Tamado 2006; Kuva et al. 2007). However, sometimes it is not possible to get proper data for such methodologies. For example, weed species identification for assessing the whole weed community identity is usually difficult, time consuming, and often requires a significant degree of expertise relating to weed taxonomy (Gonzalez-Andujar et al. 2006). Regarding the explanatory factors, the rise in the use of geographic information systems linked to cropping management increases the availability of direct field data; but data inconsistency is likely to occur (i.e., missing data or unbalanced designs) (Kenkel et al. 2002). However, a process of data mining (i.e., the nontrivial extraction of hidden and potentially useful information and significant knowledge from large sets of data) would be highly desirable to detect patterns that can be further refined by using higher statistical power approaches.

In sugarcane cropping systems the analysis of direct field information has delivered adequate information regarding trends in productivity (Ellis et al. 2001; Ferraro et al. 2009) or changes due to regional or climate variability (Lawes et al. 2004; Russell et al. 1991). In this study we proposed the use of a data-mining technique called classification and regression trees (CART) (Breiman et al. 1984) as an exploratory method for assessing the effect of environmental and management factors on the weed composition. This technique is particularly suitable for analyzing large databases covering large numbers of variables that usually involve nonlinear relationships and complex interactions (Garzón et al. 2006; Peltzer et al. 2008). Also, outputs from CART are summarized in a tree that facilitates the outcome results. Moreover, in vegetation analysis, a CART analysis could be more powerful to detect trends and patterns than traditional multivariate analysis (e.g., canonical correspondence or multivariate redundancy analyses) because it is possible not only to include continuous but also categorical variables among the explanatory variable group without any assumption of normality (Steinberg and Colla 1995). The CART analytic approach has been previously used to study seed bank dynamics (Wiles and Brodahl 2004), and to predict weed population abundances and distribution (Debeljak et al. 2008). This study uses data-mining techniques to define the main effects among the many potential interactions between management practices, environmental conditions, and the most conspicuous weed species in a large number of sugarcane fields. We proposed an exploratory data analysis of weed composition by using direct field information instead of a manipulative or controlled experimental approach. The goal of this exploratory analysis is to uncover the most important factors for predicting different weed clusters. Particularly, we studied sugarcane cropping systems of northern Argentina where weeds are a major problem and the hierarchy of factors

DOI: 10.1614/WS-D-11-00023.1

* Assistant Professor, IFEVA, Cátedra de Cerealicultura, Facultad de Agronomía, Universidad de Buenos Aires/CONICET, Av. San Martín 4453, Buenos Aires (1417DSE), Argentina; Professor and graduate student, IFEVA, Cátedra de Ecología. Facultad de Agronomía, Universidad de Buenos Aires/CONICET. Corresponding author's E-mail: ferraro@agro.edu.ar

affecting weed composition is poorly understood. The objectives of the study were to (1) characterize the clustering of the most abundant and problematic weed species in sugarcane, and (2) explore the effects of environmental and management variables for predicting the characterized clusters.

Materials and Methods

Study Area. The sugarcane fields sampled were located in Jujuy province in northern Argentina (23°N, 65°W at 670 m elevation), a relatively warm (20 to 22 C) tropical region with an annual rainfall of 800 mm received during the summer period between December and March. Sugarcane is grown under irrigation in this region. The major soil types in the study area were sandy loam (Udic Haplustalfs) in the low-landscape areas, silty loam (Udic Argiustolls) in the mid-slopes and ridge-top landscape areas, and clayey loam (Typic Ustifluvents) on alluvial areas close to riverbanks or flood plains. Until recently, sugarcane fields were burned before harvesting, but this practice is being replaced by green-cane harvesting. In the green-cane system, harvesters return crop residues (leaves and tops) to the ground surface, which results in both substantial improvements in profitability through cost savings and in soil organic matter content, nutrient and water retention, and soil biodiversity (Basanta et al. 2003; Braunbeck et al. 1999; Garside et al. 1997; Vallis et al. 1996; Wood 1991). This trash blanket (i.e., retention of the residues as a surface mulch), along with the lack of disturbance of row tops, often leads to the proliferation of perennial weeds in the sugarcane crop (Richard 1995).

Sampling. As a joint effort by local producers to monitor the level of weed infestation in their sugarcane crops, data on abundance of the most important weed species, in terms of difficulty of control, frequency, and potential crop yield loss (Table 1), were recorded by local growers during 2004 and 2005. CART models require a large number of observations to get a reliable assessment. Thus, we used the grower's records for building a large database of weed composition and explanatory factors from 1976 sugarcane fields in the study area. The database structure had no evident bias, as growers randomly selected the sugarcane fields for weed sampling. Weed surveys were performed by two or more trained people who walked across each field recording the predefined list of species. Sampling was restricted to areas in the field with homogeneous crop cover by avoiding field margins. Abundance of each species was estimated by considering the percentage of ground cover using the following estimate considering the percentage of ground cover, with the following scale: 0 to 1 (rare = 0), 2 to 10 (low = 1), 11 to 30 (medium = 2), 31 to 60 (high = 3), 61 to 100% (very high = 4) (Mueller-Dombois and Ellenberg 1974). The analyzed weed abundance database included two consecutive harvest seasons (2004 and 2005) in ratoon-cane stands and was collected from August to January. All weed surveys were done during the tillering phase before the stalk growth period when the leaf area index reached the maximum value. Moreover, we included both harvest and sampling time as explanatory variable (see below) to quantify the effect of sampling time on the composition of the weed species recorded.

Table 1. Weed species recorded for data analyses in sugarcane fields.

Species	Abbreviation[a]
Annuals	
Setaria viridis (L.) Beauv.	SETVI
Rottboellia exaltata L.	ROOEX
Digitaria sanguinalis (L.) Scop	DIGSA
Polygonum convolvulus L.	POLCO
Leptochloa filiformis (Lam.) P. Beauv	LEFFI
Tithonia tubaeformis (Jacq.) Cass.	TITTU
Trianthema portulacastrum L.	TRTPO
Other broadleaf species	BBBAN
Perennials	
Cynodon dactylon (L) Pers.	CYNDA
Cyperus rotundus L.	CYPRO
Sorghum halepense (L.) Pers	SORHA

[a] Letter code for weed names in Weed Science Society of America-approved computer code from the *Composite List of Weeds*. Available at http://wssa.net/Weeds/ID/WeedNames/namesearch.php

Cluster Analysis. Variability in weed composition was assessed using Bray–Curtis distance measure (McCune and Mefford 1995). The abundance matrix was used for calculating a dissimilarity matrix (Shaw 2003) for each of the 1976 samples. The dissimilarity data were clustered through a *k*-means cluster algorithm (Jain and Dubes 1988) to detect different weed groups and the membership of each sugarcane field to one of them. The clustering algorithm is based on a least sum-of-squares estimation, and attempts to group the sugarcane fields by reducing the intragroup variability, in terms of floristic composition, as well as maximize the intragroup variability. A standard cross-validation procedure was applied for determining the final number of weed clusters (*n*). In cross-validation, the total number of samples is divided into v "folds" samples of approximately equal size. Then, cluster analyses for each v − 1 samples (i.e., leaving out one sample) are repeated, and the samples that were not used to compute the respective cluster solution are treated as test samples, for which the average distance of observations from their respective (assigned) cluster centers will be recorded. This measure is the misclassification error (*r*) (i.e., the proportion of data points assigned to an incorrect cluster) and the procedure is repeated by increasing the number of tested clusters. Finally, for selecting the optimal number of weed clusters we inspected the set of solutions for detecting: (1) a cutoff value of 5% in the percentage decrease in the misclassification error when adding one more cluster and (2) the lowest number of clusters that meets the above condition. Both conditions determine a proper balance between accuracy and complexity for the final number of clusters (Jain 2010; Koziol 1990). After selecting the final cluster configuration, the Shannon's diversity index (*H*) (Magurran 1988) was calculated for each cluster:

$$\text{Shannon's diversity index}(H') = -\sum p_i \ln p_i \qquad [1]$$

where p_i is the proportion of the weed abundance of i species in the total sample of species in each cluster.

Classification and Regression Tree Analysis. CART was used for partitioning the clustered weed groups into subsets (or nodes) with the highest attainable homogeneity defined by several explanatory factors (Tables 2 and 3). Basically, a classification tree partitions the space of all possible attributes (both

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press

Table 2. Description of continuous sugarcane management variables used for the study.

Variable	Abbreviation	Unit	Cases	Mean	Max	Min	25th–75th percentile
Field area	AREA	ha	1,976	40.02	157.88	0.75	24–72
Crop class	AGE	yr	1,976	2.40	19	1	1–3
Herbicide applications	HERB	number of applications	1,976	5.42	20	1	3–7
Sugarcane biomass yield	YIELD	t ha^{-1}	777	99.2	294	1.8	85–114
Interval between sampling and harvest	SHI	d	1,906	80.5	207	2	50–107
Annual precipitation	PP	mm	1,976	703	1,020	420	632–777

categorical and continuous), starting with all attributes (at the root of the tree) and successively splitting that space in nodes in which each node is more likely to be assigned to one of the *k*-means clusters than the node from which it is split (Breiman et al. 1984). Ideally, the process of splitting continues until each node is pure (i.e., contains only one class of elements) or the gain in purity of the final nodes (i.e., terminal nodes) reaches a certain threshold. CART models are also extremely robust on the effects of outliers as well as being able to deal with missing values by minimizing or eliminating the effect of such values on model performance. In this study, CART algorithm was used to generate a threshold to differentiate nodes and to find a tree structure that discriminates the *k*-means clusters (i.e., which terminal nodes have a high proportion of sugarcane fields of some cluster). A standard cross-validation (CV) procedure was applied for measuring the predictive power of the trees obtained (Waheed et al. 2006). The main idea of cross-validation is that each observation is included in both the test sample and the training sample. We used the standard cross-validation, where a data set is randomly divided into 10 parts. Iteratively, 10 different models are generated, each iteration involving a different combination of nine parts for model development (training or learning) and one part for testing (cross-validation). The tree-growing process is repeated 10 times, and when completed the error counts from each of the 10 test samples are summed to obtain the CV error estimate (i.e., the proportion of cases incorrectly classified in the tree). In CART the equivalent to the R^2 of linear regression is (1 − CV error) (Breiman et al. 1984). This estimates the "portion of variance explained by the model" (Roel et al. 2007). Finally, the CART procedure considers the importance of the independent variables, which are ranked in descending order of their contribution to tree construction. This contribution is not necessarily associated to the relative position of the variable in the tree structure, because the procedure looks at the improvement measure attributable to each variable in its role as a surrogate to the primary split in each node splitting. The values of these improvements are summed over each node of the tree and scaled relative to the best-performing variable. The variable with the highest sum of improvements is scored 1 and all other variables have lower scores ranging downward toward zero (Steinberg and Colla 1995).

Explanatory Factors. Agronomic and environmental data used as explanatory factors for weed composition are listed in Tables 2 and 3: field area (AREA); crop class (AGE); number of herbicides applied during the growth cycle (HERB); sugarcane biomass harvested yield (YIELD); time interval between sampling and harvest (SHI); time (month) of sample (SAMPLE); amount of rain during the 12 mo before weed survey (PP); sugarcane genotype (GEN); soil quality (SOIL); and crop field with burning or not burning of crop stubble (BURN). Field area (an indirect indicator of the relative importance of field borders) was used to detect landscape or

Table 3. Description of categorical sugarcane management variables used for this study.

Variable	Abbreviation	Code	Cases	Area (ha)	Factor's levels
Sugarcane cultivar	GEN	1	33	1,239	CP 65–350
		2	34	1,897	CP 68–350
		3	158	7,993	CP 70–1133
		4	248	8,928	CP 72–2086
		5	113	4,715	NA 84–3920
		6	525	20,033	NA 85–1602
		7	137	5,289	TUC 67–24
		8	134	5,853	TUC 72–16
		9	572	22,328	TUC 77–42
		0	22	808	Other
Soil quality	SOIL	1	50	2,923	A1
		2	215	8,132	A2
		3	703	25,697	A3
		4	469	18,429	A4
		5	180	8,029	A5
		0	359	15,876	No data
Burned harvest	BURN	0	1376	53,526	Green harvest
		1	600	25,559	Burned harvest
Month of sample	SAMPLE	8	36	1,234	August
		9	269	10,697	September
		10	557	21,790	October
		11	760	31,089	November
		12	343	13,720	December
		13	11	551	January

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press

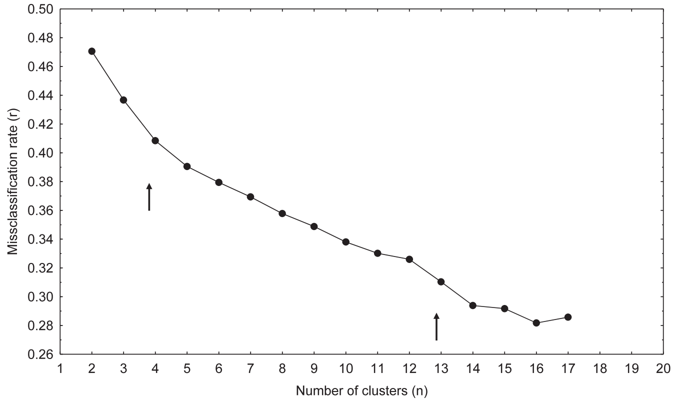

Figure 1. Misclassification rate (*r*) of sugarcane fields used as test set during the cross-validation procedure for optimizing the final number of clusters (*n*). The error function (*r*) is calculated as the average distance of sugarcane fields used as test samples to the cluster centroids to which they were assigned. The left and right arrows indicate the first and last cluster number where the percentage of *r* decrease is less than 5%, respectively.

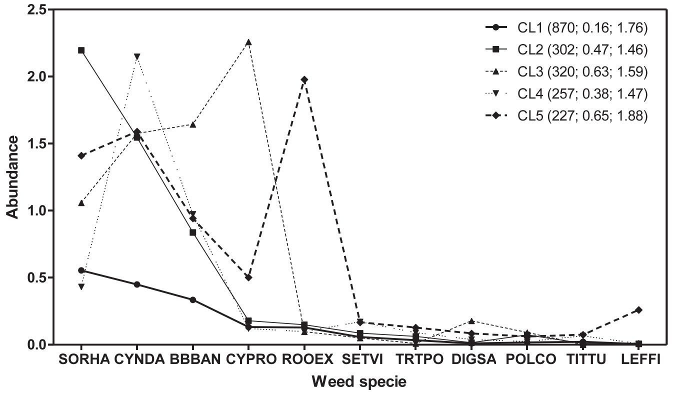

Figure 2. Mean of weed species abundance in each of the five clusters (CL) selected. Numbers between brackets indicate the number of cases, the overall mean abundance, and the Shannon's diversity index (*H*′) for each cluster, respectively. The abbreviations of weed species are listed in Table 1.

geographical effects on weed composition (i.e., weed dispersal). Crop class indicated both the number of ratoon crops and the years after the plant-cane crop, because the sugarcane crops have one harvest per year. Month of sampling (SAMPLE) was restricted to the tillering phase before the stalk growth period (August to January), but most of the cases fell between September and the end of December. Annual precipitation was estimated using data of the previous 365 d after the weed survey from the closest weather station to each field. Nine sugarcane cultivars were included in the study, covering a wide range of biomass yield potentials and sucrose accumulation dynamics (Table 3). Soil quality was characterized with the widely used land capability class index that ranges from 1 (highest quality) to 8 (lowest quality), which is based on a ranking of 12 different soil characteristics that are critical for crop production (USDA 1989). Class 1 has no significant limitations for raising crops. Classes 2 and 3 are suited for cultivated crops but have limitations such as poor drainage, limited root zones, climatic restrictions, or erosion potential. Class 4 is suitable for crops but only under selected cropping practices and class 5 is best suited for pasture and range.

Results and Discussion

Cluster Analysis. The field misclassification rate (i.e., the proportion of errors made on the classification procedure) decreased until the databases were divided into $n = 16$ individual clusters (Figure 1). However, the 5% cutoff value of rate reduction was reached at the $n = 5$ and $n = 14$ levels of cluster splitting. Therefore, the $n = 5$ cluster number was selected for further analyses. This final cluster configuration showed one main cluster (CL1) that contained 44% of the sugarcane cases analyzed and also exhibited the lowest mean abundance value (Figure 2). The other four clusters were dominated by the perennial species bermudagrass, johnsongrass, and purple nutsedge and the annual itchgrass (Figure 2). Both the overall abundance and the species diversity contributed to the dissimilarity among clusters (Figure 2). CL2 and CL4 were dominated by johnsongrass and bermudagrass, and exhibited low values of both overall weed abundance and diversity. In contrast, the cluster dominated by the annual itchgrass (CL5) was the most diverse weed cluster showing mean abundance values higher or equal to 1.0 for four weed species (johnsongrass, bermudagrass, other broad-leaf [BBBAN], and itchgrass). Weed composition described in this study was similar to other sugarcane regions where annual weeds are the most common (in the absence of any control measure), but perennial weeds are among the most difficult to control (Kuva et al. 1999; McMahon 1989; Smith 1998). Perennials have propagules such as rhizomes, bulbs, or tubers that are difficult to eliminate and contribute to weed persistence and dispersal and increase their aggressiveness (Ali et al. 1986; Bariuan et al. 1999; Holm et al. 1977). In sugarcane, perennial infestation may be especially enhanced because the row top is not disturbed over the 3- to 5-yr crop cycle and for this reason weed management practices are focused on reducing perennial infestation during plant-cane stage (Peng 1984).

Classification and Regression Tree Analysis. Final CART model was able to explain 44% of the sugarcane's weed composition, a relatively high value for vegetation analysis (Ter Braak and Prentice 1988). Four of the five clusters were represented in the terminal nodes of the final CART model (Figure 3). The exception was CL5, the most diverse of the five clusters. CART started the splitting process by dividing the root node (Figure 3, $\text{ID} = 1$) into two subgroups. Sugarcane fields where the sugarcane trash was burned or not burned were split on the left or right main branches of the classification tree, respectively. The left branch of the tree showed two more divisions, resulting in three terminal nodes that were classified as CL1 cluster (the more frequent cluster with the lowest abundance value). Although the detrimental effects of burning cane residue on nutrient availability due to the loss of organic matter can be thought of as short term, results of this study indicated that the effects would extend beyond 70 d after harvest (Figure 3, $\text{ID} = 5$). Moreover, under these conditions, when annual amount of rain was higher than 607 mm, CART was able to split low-abundance sugarcane fields with high accuracy (76% of the sugarcane fields were in agreement with the selected terminal node classification) (Figure 3; $\text{ID} = 7$). Research on the effect of burning on weed community composition is scarce, but (Galdos et al. 2010) noted more growth of weeds in burned plots, probably due to higher competition with the sugarcane

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press

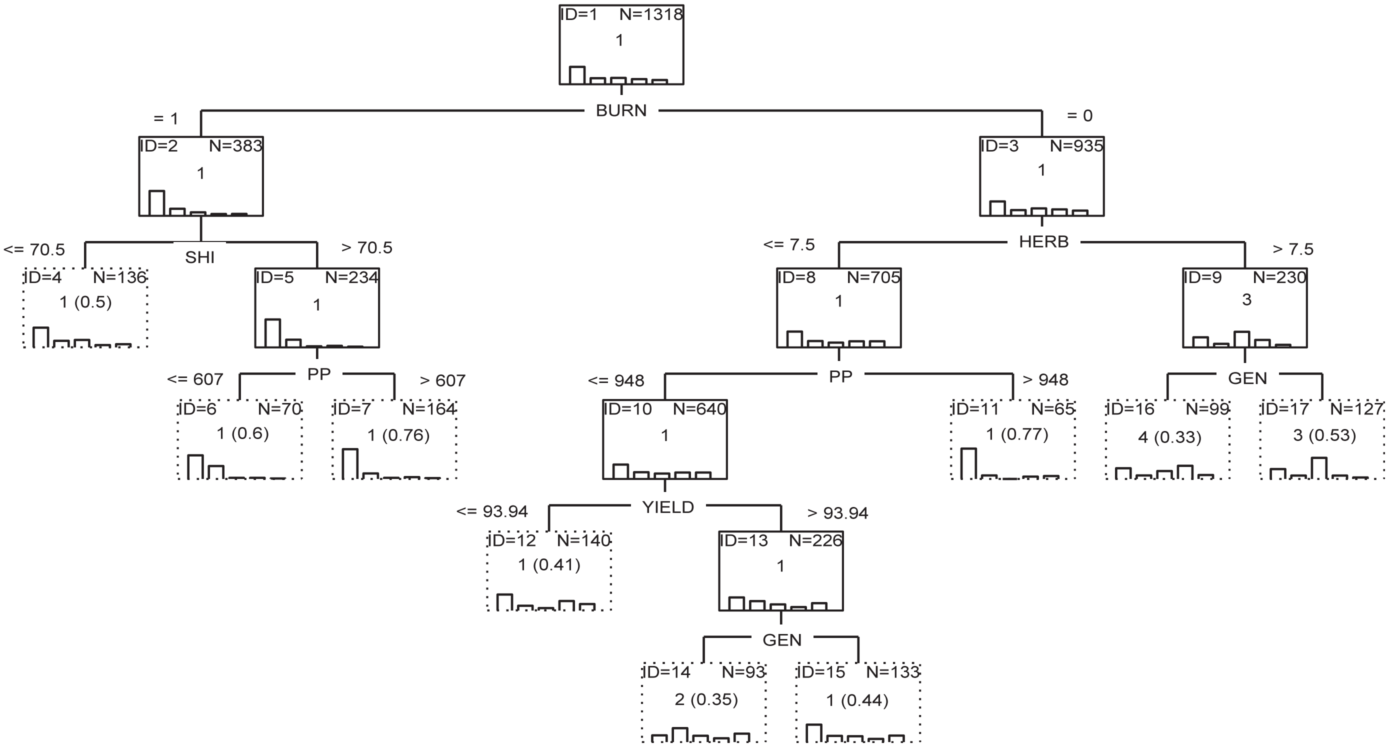

Figure 3. Classification tree (CART) model of sugarcane crops, using the *k*-means cluster identified in Figure 2. *N* indicates the number of sugarcane fields in that node of the classification tree. Right and left branches indicate that the group satisfies, or does not satisfy, respectively, the split condition at a decision node. For variable abbreviations see Tables 2 and 3. Dotted boxes indicate terminal nodes. Columns inside each node are the cluster distribution histogram, and the number in the center, the most frequent cluster (with its frequency between brackets). Misclassification error (r) for the learning set (1,318 cases) = 0.56. Misclassification error (r) for test set (658 cases) = 0.52 (no significant difference, $P < 0.05$). Variance explained by the model: 1. Misclassification error (r) = 0.44.

for water and nutrients, and the lower efficiency of herbicide application in the mulch retention system (Christoffoleti et al. 2007). In contrast, results from this study showed that CART associated burned sugarcane fields with relatively low values of weed abundance (CL1). The emerging relationship between preharvest burning and low abundance of weeds would indicate a rapid growth fueled by intensive resource uptake (light) that allows the domination of sugarcane in the process of crop–weed competition. This means that this variable would affect the weed presence/absence more than the identity of the species assemblage. However, the use of a more accurate multivariate method (canonical correlation analysis) in a more controlled condition (using the full species list) could find some composition effect between green and burned harvest.

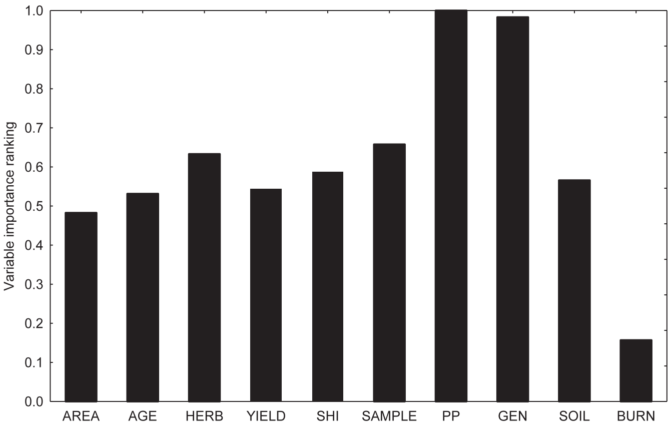

Figure 4. Variable importance ranking computed by the CART model of Figure 3. The abbreviations of weed species are listed in Table 1.

The right split from the root node isolated the unburned sugarcane fields (Figure 3, ID = 3), and progressively split the data set into smaller subsets until six terminal nodes were obtained (Figure 3). Three of these terminal nodes classified as CL1, and came from an intermediate split that selected the use of herbicides as the partition variable (Figure 3, ID = 3). Values of herbicide use lower than 7.5 herbicide/cropping cycle resulted in the final terminal nodes for CL1, on the basis of an interaction with annual precipitation level (Figure 3, ID = 11). This pattern was similar to that observed in the unburned sugarcane fields of the left side of the tree, where higher precipitation was more associated with sugarcane fields with low values of weed abundance (Figure 3, ID = 7). However, there was a significant difference in the cutoff value for precipitation in each split node (607 mm yr^{-1} for burned fields and 948 mm yr^{-1} for unburned fields). Both values are slightly below and above 800 mm of annual average precipitation. Differences in effective precipitation due to the mulch presence associated with the green harvest system could explain the higher cutoff values of precipitation in green harvest conditions than in preharvest burning for this splitting step. Also, CL1 was selected in this part of the tree with annual precipitation values lower than 948 mm, but with the additional condition of crop yields lower than 93.94 t ha^{-1} (Figure 3, ID = 12) or under higher crop values but for only some select genotypes (Figure 3, ID = 15). These results showed that the CART algorithm was able to select several paths for assessing the condition of lowest weed abundance (CL1).

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press

CART structure showed the number of herbicides as a selected predictor for splitting the green-harvested sugarcane fields into two groups. Surprisingly, the most intensive herbicide use (values higher than 7.5 herbicides used in one cropping cycle) was not associated with weed cluster with low abundance values (i.e., CL1). Moreover, the cropping system with the more intensive herbicide use was associated with CL3 and CL4, two clusters dominated by the perennials purple nutsedge and bermudagrass (Figures 2 and 3). Although perennials are difficult to control with herbicides, it is not possible to determine whether the use of herbicides is a cause or a consequence of the observed group of weeds. Consequently, these results highlight the importance of assessing the relationship between the processes of weed growth and the frequency of herbicide use in the systems studied.

Another factor that was important when selecting a weed cluster for each sugarcane field was the sugarcane genotype (Figure 4). The final CART was able to show a genotype-related effect for splitting CL3 and CL4 weed groups (Figure 3, IDs = 16 and 17). Although it did not identify any genotype trait, CART would show a possible existence of direct effects (e.g., competition) or indirect (e.g., chemical rhizosphere composition) on the filtering of weed clusters. There is evidence in the literature about the genotypic effects in the variability of thermal attributes (Liu et al. 1998), growth and development of root system (Smith et al. 2005), allelopathy (Sampietro et al. 2006), and the final biomass and sugar yield (Kang et al. 1987). Clearly, this variability could affect the ability for faster canopy closing and hence reduce weed competition in the initial 90- to 120-d period of crop growth that is considered the most critical period of weed competition in sugarcane (Kuva et al. 1999).

CART and variable importance ranking highlighted data patterns not only through tree selection variables but also by the unselected variables in the final tree. For example, soil type and crop age are two important variables for defining final sugarcane yield (Evenson et al. 1987; Muchow et al. 1996), but they were not selected in final CART configuration, and exhibited intermediate variable importance ranking. Usually, crop age is related to yield decline in sugarcane because of monoculture, excessive tillage, and the decline of soil health (Magarey et al. 1997; Pankhurst et al. 2003; Pankhurst et al. 2005). However, results of this study would indicate that both plant and soil changes related to sugarcane ratooning did not have a clear and detectable effect on weed heterogeneity among the large database of sugarcane fields analyzed. Probably the more evident change in weed composition would be observed between plant and ratoon cane, so the crop age effect would be qualitative rather than quantitative. Although the soil type was not able to explain differences in weed composition, weed dynamics process could be spatially autocorrelated because of altitudinal changes or physical dispersal constraints (Booth and Swanton 2002). The studied fields were distributed among very different landscapes that included valleys, piedmonts, river terraces, flood plains, and alluvial plains. Some of these soils are prone to water erosion and irrigation management difficulties. For example, piedmonts and alluvial plains are moderately to marginally suitable for sugarcane production, especially because of poor soil water retention, topography, or the presence of a high water table. However, because of this large soil gradient the abiotic conditions reflected by soil did not appear to be strong enough to be selected in the CART structure. It is also possible that the weed community characterization using the predefined list of the most conspicuous species would hinder the detection of weaker effects (as might be the soil type), since these species would most likely have to be in the range of sites surveyed. The rest of the predictors showed importance values in the range of 0.48 to 0.62 that correspond to the proportion of the variance explained by the most powerful splitting variable (PP) (Figure 4).

Results from this study identified homogeneous groups in the composition of the most important weeds in a sugarcane agroecosystem through the characterization of hierarchical models that explained weed composition and the overall importance of each predictor. The analytical framework (data mining) is presented as a possible protocol for studying weed composition in crops using incomplete data. Through robust analyses, data mining was able to obtain exploratory patterns in the crop–weed system to be further refined under more controlled conditions (manipulative experiments or census of all species in the community). Although its application in this study was done on sugarcane, it is clear that it could extend to any other system where there can be available management and environmental databases, for example to eliminate low importance values and help the next analysis to fine-tune its approach and reduce the data set examined. For the systems analyzed, the results showed that burned harvest was the most important factor related to low weed abundance, and that the effect was maintained over time. Regarding the predictive power of the variables, rainfall and genotype identity were the most important predictors. These results have management implications as they indicate that the genotype identity would be a more important factor than crop age when designing sugarcane weed management. Moreover, the abiotic control of crop–weed interaction would be more related to rainfall than the environmental heterogeneity related to soil type, for example soil fertility. Although all these exploratory patterns resulting from the CART data-mining procedure should be refined, it became clear that this information may be used to develop an experimental framework to study the factors driving weed assembly.

Acknowledgments

The authors thank all the personnel from the Ledesma Mill in Jujuy involved in this study for their patience, collaboration, and for allowing sampling within farms.

Literature Cited

Ali, A. D., T. E. Reagan, L. M. Kitchen, and J. L. Flynn. 1986. Effects of johnsongrass (*Sorghum halepense*) density on sugarcane *(Saccharum officinarum*) yield. Weed Sci. 34:381–383.

Bariuan, J. V., K. N. Reddy, and G. D. Wills. 1999. Glyphosate injury, rainfastness, absorption, and translocation in purple nutsedge (*Cyperus rotundus*). Weed Technol. 13:112–119.

Basanta, M. V., D. Dourado-Neto, and K. Reichardt, et al. 2003. Management effects on nitrogen recovery in a sugarcane crop grown in Brazil. Geoderma 116:235–248.

Booth, B. D. and C. J. Swanton. 2002. Assembly theory applied to weed communities. Weed Sci. 50:2–13.

Braunbeck, O., A. Bauen, F. Rosillo-Calle, and L. Cortez. 1999. Prospects for green cane harvesting and cane residue use in Brazil. Biomass Bioenergy 17:495–506.

Breiman, L., R. Friedman, R. Olshen, and C. Stone. 1984. Classification and Regression Trees. Boca Raton, FL: CRC Press. 368 p.

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press

Christoffoleti, P. J., S.J.P. de Carvalho, R. F. López-Ovejero, M. Nicolai, E. Hidalgo, and J. E. da Silva. 2007. Conservation of natural resources in Brazilian agriculture: implications on weed biology and management. Crop Prot. 26:383–389.

Debeljak, M., G. R. Squire, D. Demsar, M. W. Young, and S. Dzeroski. 2008. Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the UK. Ecol. Model. 212:138–146.

Ellis, R. N., K. E. Basford, M. Cooper, J. K. Leslie, and D. E. Byth. 2001. A methodology for analysis of sugarcane productivity trends. I. Analysis across districts. Aust. J. Agric. Res. 52:1001–1009.

Evenson, C. I., R. C. Muchow, S. A. El-Swaify, and R. V. Osgood. 1987. Yield accumulation in irrigated sugarcane. I. Effect of crop age and cultivar. Agron. J. 89:638–646.

Ferraro, D. O., D. E. Rivero, and C. M. Ghersa. 2009. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. Field Crop. Res. 112:149–157.

Firehun, Y. and T. Tamado. 2006. Weed flora in the Rift Valley sugarcane plantations of Ethiopia as influenced by soil types and agronomic practises. Weed Biol. Manag. 6:139–150.

Galdos, M., C. Cerri, C. Cerri, K. Paustian, and R. Van Antwerpen. 2010. Simulation of sugarcane residue decomposition and aboveground growth. Plant Soil 326:243–259.

Garside, A. L., M. A. Smith, L. S. Chapman, A. P. Hurney, and R. C. Magarey. 1997. The yield plateau in the Australian sugar industry: 1970–1990. Pages 103–124 *in* B. A. Keating and J. R. Wilson, eds. Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. Wallingford, UK: CAB International.

Garzón, M. B., R. Blazek, M. Neteler, R.S.d. Dios, H. S. Ollero, and C. Furlanello. 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecol. Model. 197:383–393.

Gonzalez-Andujar, J. L., C. Fernandez-Quintanilla, J. Izquierdo, and J. M. Urbano. 2006. SIMCE: an expert system for seedling weed identification in cereals. Comp. Electron. Agr. 54:115–123.

Holm, L. G., D. L. Plucknett, J. V. Pancho, and J. P. Herberger. 1977. The World's Worst Weeds: Distribution and Biology. Honolulu, HI: University Press of Hawaii. 609 p.

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31:651–666.

Jain, A. K. and R. C. Dubes. 1988. Algorithms for Clustering Data. New Jersey: Prentice-Hall. 320 p.

Kang, M. S., J. D. Miller, P.Y.P. Tai, J. L. Dean, and B. Glaz. 1987. Implications of confounding of genotype × year and genotype × crop effects in sugarcane. Field Crop. Res. 15:349–355.

Kenkel, N. C., D. A. Derksen, A. G. Thomas, and P. R. Watson. 2002. Multivariate analysis in weed science research. Weed Sci. 50:281–292.

Koziol, J.A.G. 1990. Cluster analysis of antigenic profiles of tumours: selection of number of clusters using Akaike's information criterion. Method. Inform. Med. 29:200–204.

Kuva, M., P. Christoffoleti, and P. Pitelli. 1999. Critical period of competition between sugarcane and weeds in Brazil. Weed Sci.Soc. Am. Abstr. 25 p.

Kuva, M. A., R. A. Pitelli, T. P. Salgado, and P.L.C.A. Alaves. 2007. Fitossociologia de comunidades de plantas daninhas em agroecossistema cana-crua. Planta Daninha 25:501–511.

Lawes, R. A., R. J. Lawn, M. K. Wegener, and K. E. Basford. 2004. The evaluation of the spatial and temporal stability of sugarcane farm performance based on yield and commercial cane sugar. Aust. J. Agric. Res. 55:335–344.

Liu, D. L., G. Kingston, and T. A. Bull. 1998. A new technique for determining the thermal parameters of phenological development in sugarcane, including suboptimum and supra-optimum temperature regimes. Agr. Forest Meteorol. 90:119–139.

Magarey, R. C., H. Y. Yip, J. I. Bull, and E. J. Johnson. 1997. Effect of the fungicide mancozeb on fungi associated with sugarcane yield decline in Queensland. Mycol. Res. 101:858–862.

Magurran, A. E. 1988. Ecological Diversity and its Measurement. London: Croom Helm. 179 p.

Martínez-Ghersa, M. A., C. M. Ghersa, and E. H. Satorre. 2000. Coevolution of agricultural systems and their weed companions: implications for research. Field Crop. Res. 67:181–190.

McCune, B. and M. J. Mefford. 1995. PC-ORD: multivariate analysis of ecological data. Gleneden Beach, OR: MjM Software Design.

McMahon, G. 1989. Weeds reduce cane yield in early growth stages. Brisbane, Australia: Bureau of Sugar Experiment Station. Sugar Exp. Stn. Bull., 27 (July). pages 21–32

Muchow, R. C., M. J. Robertson, and A. W. Wood. 1996. Growth of sugarcane under high input conditions in tropical Australia. II. Sucrose accumulation and commercial yield. Field Crop. Res. 48:27–36.

Mueller-Dombois, D. and H. Ellenberg. 1974. Causal analytical inquiries into the origin of plant communities. Pages 335–370 *in* Aims and Methods of Vegetation Ecology. New York: Wiley.

Pankhurst, C. E., R. C. Magarey, G. R. Stirling, B. L. Blair, M. J. Bell, and A. L. Garside. 2003. Management practices to improve soil health and reduce the effects of detrimental soil biota associated with yield decline of sugarcane in Queensland, Australia. Soil Till. Res. 72:125.

Pankhurst, C. E., G. R. Stirling, R. C. Magarey, B. L. Blair, J. A. Holt, M. J. Bell, and A. L. Garside. 2005. Quantification of the effects of rotation breaks on soil biological properties and their impact on yield decline in sugarcane. Soil Biol. Biochem. 37:1121–1130.

Peltzer, D. A., S. Ferriss, and R. G. FitzJohn. 2008. Predicting weed distribution at the landscape scale: using naturalized *Brassica* as a model system. J. Appl. Ecol. 45:467–475.

Peng, S. Y. 1984. The Biology and Control of Weeds in Sugarcane. New York: Elsevier Science. 336 p.

Richard, E. P., Jr. 1995. Bermudagrass interference during a three year sugarcane crop cycle. Proc. Int. Soc. Sugar Cane Technol. 21:31–39.

Roel, A., H. Firpo, and R. E. Plant. 2007. Why do some farmers get higher yields? Multivariate analysis of a group of Uruguayan rice farmers. Comp. Electron. Agric. 58:78–92.

Russell, J. S., M. K. Wegener, and T. R. Valentine. 1991. Effect of weather variables on C.C.S. at Tully simulated by the AUSCANE model. Proc. Aust. Soc. Sugar Cane Technol. 13:157–163.

Sampietro, D. A., M. A. Vattuone, and M. I. Isla. 2006. Plant growth inhibitors isolated from sugarcane (*Saccharum officinarum*) straw. J. Plant Physiol. 163:837–846.

Shaw, P. J. 2003. Multivariate Statistics for the Environmental Sciences. New York. 233 p.

Smith, D. M., N. G. Inman-Bamber, and P. J. Thorburn. 2005. Growth and function of the sugarcane root system. Field Crop. Res. 92:169–183.

Smith, D. T. 1998. Weed Control in US Sugarcane. Technical Report 98-03. Texas: USDA Department of Soil and Crop Science, CollegeStation, TX: TexaS A&M University. Rep. 98-03. 25 p.

Steinberg, D. and P. Colla. 1995. CART: Tree-Structured Non-Parametric Data Analysis. San Diego, CA: Salford Systems. 336 p.

Ter Braak, C.J.F. and C. Prentice. 1988. A theory of gradient analysis. Adv. Ecol. Res. 18:271–317.

[USDA] U.S. Department of Agriculture. 1989, The Second RCA Appraisal: Soil, Water and Related Resources on Nonfederal Land in the United States.: U.S. Department of Agriculture, Soil Conservation Service., 280 p.

Vallis, I., W. J. Parton, B. A. Keating, and A. W. Wood. 1996. Simulation of the effects of trash and N fertilizer management on soil organic matter levels and yields of sugarcane. Soil Till. Res. 38:115–132.

Waheed, T., R. B. Bonnell, S. O. Prasher, and E. Paulet. 2006. Measuring performance in precision agriculture: CART—a decision tree approach. Agric. Water Manag. 84:173–185.

Wiles, L. and M. Brodahl. 2004. Exploratory data analysis to identify factors influencing spatial distributions of weed seed banks. Weed Sci. 52:936–947.

Wood, W. 1991. Management of crop residues following green harvesting of sugarcane in North Queensland. Soil Till. Res. 20:69–85.

Received February 22, 2011, and approved August 8, 2011.

https://doi.org/10.1614/WS-D-11-00023.1 Published online by Cambridge University Press